[Bank Marketing Data Set](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing)

By Praneeth Vepakomma

Phase -1 : Download, load data into ipython notebook (jupyter), EDA, Checks

* Checks:
  + Is there any missing data
  + **Which columns are numeric**

**Question: Can/should we write code to answer this question. Do you have code snippet for it?**

* + **Which columns are categorical**

**Question: Can/should we write code to answer this question. Do you have code snippet for it?**

* + Output: Binary classification / multi-class?
  + **Are there columns with zero or near-zero variance**
  + **Are labels/class unbalanced or balanced**

Phase-2: Preprocessing

* Only for numerical cols: apply centering and scaling(normalization) operations
  + <https://scikit-learn.org/stable/modules/preprocessing.html#standardization-or-mean-removal-and-variance-scaling>
* Only for categorical cols: Apply one-hot encoding (also called dummy coding)
  + <https://scikit-learn.org/stable/modules/preprocessing.html#encoding-categorical-features>
* Are there columns with zero- if so , just drop those columns
  + or near-zero variance-examine if it is abnormal or useless variable depending on your project situation-drop or don't drop accordingly
* Are labels/class unbalanced or balanced
  + Get counts (and proportions) of each unique label type
  + If imbalanced-apply SMOTE sampling to generate synthetic data to append to original dataset-such that the newer dataset is more balanced
    - <https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html>

Phase - 3: Initial model

Phase- 4: Model tuning and evaluation

Phase-5: Reporting